

IN THE CLAIMS

Please amend or retain the claims as follows:

1. (Currently Amended) A method of setting up a DLSI space-based classifier to be stored in a computer storage device for document classification using a computer, and classifying using said classifier by said computer to classify a document according to a plurality of clusters within a database, using said classifier comprising the steps of:

preprocessing documents using said computer to distinguish terms of a word and a noun phrase from stop words;

constructing system terms by setting up a term list as well as global weights using said computer;

normalizing document vectors of collected documents, as well as centroid vectors of each cluster using said computer;

constructing a differential term by intra-document matrix $D_I^{m \times n_I}$ using said computer, such that each column in said matrix is a differential intra-document vector;

decomposing the differential term by intra-document matrix D_I , by an SVD algorithm using said computer, into $D_I = U_I S_I V_I^T$ ($S_I = \text{diag}(\delta_{I,1}, \delta_{I,2}, \dots)$), followed by a composition of $D_{I,k_I} = U_{k_I} S_{k_I} V_{k_I}^T$ giving an approximate D_I in terms of an appropriate k_I ;

setting up a likelihood function of intra-differential document vector using said computer;

constructing a term by extra-document matrix $D_E^{m \times n_E}$ using said computer, such that each column of said extra-document matrix is an extra-differential document vector;

decomposing D_E , by exploiting the SVD algorithm using said computer, into

$D_E = U_E S_E V_E^T$ ($S_E = \text{diag}(\delta_{E,1}, \delta_{E,2}, \dots)$), then with a proper k_E , defining $D_{E,k_E} = U_{k_E} S_{k_E} V_{k_E}^T$ to approximate D_E ;

setting up a likelihood function of extra-differential document vector using said computer;

setting up a posteriori function using said computer; and

~~classifying, on the basis of the said computer using said DLSI space-based classifier, as set up in the foregoing steps, to classify the document as belonging to one of the plurality of clusters within the database.~~

2. (Currently Amended) An automatic document classification method using a DLSI space-based classifier operating on a computer as a computerized classifier to classify a document in accordance with clusters in a database, comprising the steps of :

- a) setting up, by said computerized classifier, a document vector by generating terms as well as frequencies of occurrence of said terms in the document, so that a normalized document vector N is obtained for the document;
- b) constructing, by said computerized classifier, using the document to be classified, a differential document vector $x = N - C$, where C is the normalized vector giving a center or centroid of a cluster;
- c) calculating an intra-document likelihood function $P(x | D_I)$ for the document using said computerized classifier;
- d) calculating an extra-document likelihood function $P(x | D_E)$ for the document using said computerized classifier;
- e) calculating a Bayesian posteriori probability function using said computerized classifier $P(D_I | x)$;
- f) repeating, for each of the clusters of the data base, steps b-e;

- g) selecting, by said computerized classifier, a cluster having a largest $P(D_i | x)$ as the cluster to which the document most likely belongs; and
- h) classifying, by said computerized classifier, the document in the selected cluster within said database thereby categorizing said document for an automated document retrieval system.

3. (Original) The method as set forth in claim 2, wherein the normalized document vector N is obtained using an equation, $b_{ij} = a_{ij} / \sqrt{\sum_{k=1}^m a_{kj}^2}$.

4. (Currently Amended) A method of setting up a DLSI space-based classifier to be stored in a computer storage device for document classification using a computer, and classifying using said classifier by said computer to classify a document according to a plurality of clusters within a database, using said classifier comprising the steps of:

setting up, using said computer, a differential term by intra-document matrix where each column of the matrix denotes a difference between a document and a centroid of a cluster to which the document belongs;

decomposing, using said computer, the differential term by intra-document matrix by an SVD algorithm to identify an intra-DLSI space;

setting up, using said computer, a probability function for a differential document vector being a differential intra-document vector;

calculating, using said computer, the probability function according to projection and distance from the differential document vector to the intra-DLSI space;

setting up, using said computer, a differential term by extra-document matrix where each column of the matrix denotes a differential document vector between a document vector and a centroid vector of a cluster which does not include the document;

decomposing, using said computer, the differential term by extra-document matrix by an SVD algorithm to identify an extra-DLSI space;

setting up, using said computer, a probability function for a differential document vector being a differential extra-document vector;

setting up, using said computer, a posteriori likelihood function using the differential intra-document and differential extra-document vectors to provide a most probable similarity measure of a document belonging to a given cluster; and

classifying, said computer using said DLSI space-based classifier, as set up in the foregoing steps, to classify the document as belonging to one of the plurality of clusters within the database.

5. (Original) The method as set forth in claim 4, wherein the step of setting up a probability function for a differential document vector being a differential intra-document vector is performed using an equation,

$$P(x | D_I) = \frac{n_I^{1/2} \exp\left(-\frac{n_I}{2} \sum_{i=1}^{k_I} \frac{y_i^2}{\delta_{I,i}^2}\right) \cdot \exp\left(-\frac{n_I \varepsilon^2(x)}{2 \rho_I}\right)}{(2\pi)^{n_I/2} \prod_{i=1}^{k_I} \delta_{I,i} \cdot \rho^{(r_I - k_I)/2}},$$

where $y = U_{k_I}^T x$, $\varepsilon^2(x) = \|x\|^2 - \sum_{i=1}^{k_I} y_i^2$, $\rho_I = \frac{1}{r_I - k_I} \sum_{i=k_I+1}^{r_I} \delta_{I,i}^2$, and r_I is the rank of matrix D_I .

6. (Original) The method as set forth in claim 4, wherein the step of setting up a probability function for a differential document vector being a differential extra-document vector is performed using an equation,

$$P(x | D_E) = \frac{n_E^{1/2} \exp\left(-\frac{n_E}{2} \sum_{i=1}^{k_E} \frac{y_i^2}{\delta_{E,i}^2}\right) \cdot \exp\left(-\frac{n_E \varepsilon^2(x)}{2 \rho_E}\right)}{(2\pi)^{n_E/2} \prod_{i=1}^{k_E} \delta_{E,i} \cdot \rho_E^{(r_E - k_E)/2}},$$

where $y = U_{k_E}^T x$, $\varepsilon^2(x) = \|x\|^2 - \sum_{i=1}^{k_E} y_i^2$, $\rho_E = \frac{1}{r_E - k_E} \sum_{i=k_E+1}^{r_E} \delta_{E,i}^2$, r_E is the rank of matrix D_E .

7. (Original) The method as set forth in claim 4, wherein the step of setting up a posteriori likelihood function is performed using an equation,

$$P(D_I | x) = \frac{P(x | D_I)P(D_I)}{P(x | D_I)P(D_I) + P(x | D_E)P(D_E)},$$

where $P(D_I)$ is set to $1/n_c$ where n_c is the number of clusters in the database and $P(D_E)$ is set to $1 - P(D_I)$.

8. (Currently Amended) The method as set forth in claim 4, the step of ~~classifying the document~~ said computer using the DLSI space-based classifier to classify the document comprising the steps of:

- a) setting up a document vector by generating terms as well as frequencies of occurrence of said terms in the document, so that a normalized document vector N is obtained for the document;
- b) constructing, using the document to be classified, a differential document vector $x = N - C$, where C is the normalized vector giving a center or centroid of a cluster;
- c) calculating an intra-document likelihood function $P(x | D_I)$ for the document;
- d) calculating an extra-document likelihood function $P(x | D_E)$ for the document;

- e) calculating a Bayesian posterior probability function $P(D_i | x)$;
- f) repeating, for each of the clusters of the data base, steps b-e;
- g) selecting a cluster having a largest $P(D_i | x)$ as the cluster to which the document most likely belongs; and
- h) classifying the document in the selected cluster.

9. (Original) The method as set forth in claim 8, wherein the normalized document vector N is obtained using an equation, $b_{ij} = a_{ij} / \sqrt{\sum_{k=1}^m a_{kj}^2}$.